Repozitorij CLARIN.SI

Katja Meden

Odsek za tehnologije znanja, Institut "Jožef Stefan",
Inštitut za novejšo zgodovino

Predavanje NUK
2023-02-22

Kazalo

O repozitoriju

Repozitorij CLARIN.SI - Osnovne informacije

- Deponiranje, dostop in arhiviranje jezikovnih virov (npr. korpusi)
- avdio posnetki, besedišča, slovarji in jezikovni modeli
- Trenutno: preko 500 jezikovnih virov in 90 jezikov
  - približno 1TB podatkov
- Zaupanja vreden repozitorij (certificiran s Core Trust Seal)
- Podatki hranjeni po načelih FAIR

Poslanstvo in politika repozitorija CLARIN.SI

Poslanstvo repozitorija

- Spodbujati raziskave na področju humanistike in družboslovja
- Raziskovalcem z enovito prijavo omogočen dostop do platforme
- Vključuje podatke za številne jezike, vendar večinoma pokriva vire in tehnologije za slovenščino, hrvaščino in srbščino
  - primarni uporabniki repozitorija CLARIN.SI raziskovalci in podjetja, ki se ukvarjajo s temi jeziki.

Politika repozitorija

- Citiranje
- Etična načela hrambe podatkov v repozitoriju
- Zagotavljanje trajnosti - trajni identifikatorji (handle)

Nacionalno pomembni korpusi

- Referenčni korpusi
  - Gigafida
- Združeni korpusi
  - Metafida
- Parlamentarni korpusi
  - siParl
  - ParlaMint
- Korpusi spletne komunikacije (JANES)
  - JANES Blog
  - JANES Forum
  - JANES Tweet
  - JANES Wiki (in drugi)
- Korpusi starejše slovenščine
  - Korpus IMP

Jezikovni modeli in programska oprema

- CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages
  - slovenščina
  - hrvaščina
  - makedonščina
  - bolgarščina
- Programska oprema za procesiranje različnih jezikov na različnih nivojih:
  - tokenizacija
  - POS (part-of-speech) označevanje
  - lematizacija
  - prepoznavanje imenskih entitet (NER)
- Jezikovni modeli, ki sestavljajo CLASSLA pipeline so deponirani v CLARIN.SI repozitorij

Slovarji in leksikoni

Slovarji

- Japanese-Slovene learner's dictionary jaSlo 3.1
- Dictionary of Slovenian Phrasemes
- Terminological dictionary of electronic smoking

Leksikoni

- Semantic lexicon of Slovene sloWNet 3.1
- Slovene sentiment lexicon JOB 1.0
- Nova Beseda Frequency Lexicon

Življenjski ciklus podatkov

Od stvaritve vnosa do objave

Oddaja vnosa

- Enotna prijava v repozitorij CLARIN.SI
- Ustvarjanje novega vnosa (v skladu z navodili)
- Oddaja vnosa v uredniški pregled

Uredniški pregled:

- Urednik vnos pregleda tehnične zahteve:
  - Opisni metapodatki
  - Format podatkov (datotek)
  - Validacija podatkov (v primeru XML, CoNNL-U datotek)
- Odločitev urednika:
  - Vnos je tehnično primeren: potrditev in objava vnosa
  - Vnos ni primeren: zavrnitev vnosa, vrnjen avtorju s seznamom potrebnih popravkov
  - Popravki, ponovna oddaja in uredniški pregled

Urejanje objavljenega vnosa

Objavljen vnos

- Podatki in metapodatki postanejo prosto dostopni
  - Omejeni vnosi: prosto dostopni zgolj metapodatki
- Metapodatki so agregirani prek protokola OAI-PMH

Brisanje vnosa

- Vsak lahko zaprosi za umik vnosa
- Metpodatki se ohranijo (PiD, razen v zelo specifičnih primerih)

Spreminjanje vnosa

- Manjše (tipkarske) napake popravimo uredniki (Pomoč uporabnikom)
- Večje spremembe zahtevajo oddajo nove različice vnosa

Nove različice vnosa

- Novo različico vnosa lahko ustvari zgolj oseba, ki je vnos oddala
- Dodajanje/spreminjanje metapodatkov in nalaganje novih podatkov (uredniški pregled)
- Stara in nova različica se avtomatsko povežeta

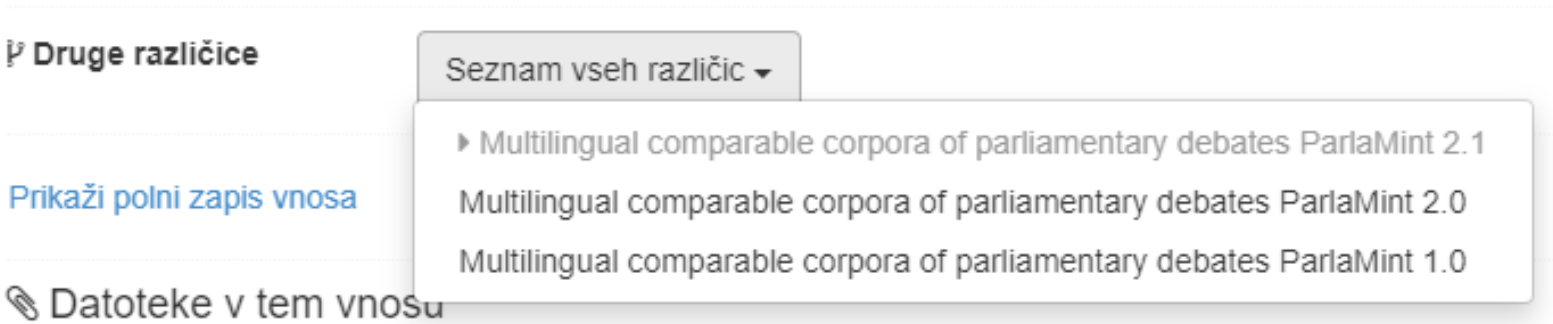


Figure: Seznam različič vnosa

Omejeni vnosi in licence

- Spodbujamo prosti dostop, vendar imajo določeni vnosi omejen dostop (omejevalne licence)
- Metapodatki omejenih vnosov so javno dostopni, za prenos podatkov je potreben el. podpis
- Vodimo evidenco elektronskih podpisov za primere kršitev pravic intelektualne lastnine

Licence

- Licence za javni dostop (Public licences):
  - Creative Commons licence
  - GNU, MIT, Apache licence
  - Ne potrebujejo nobenih dodatnih informacij
- Licence z omejenim dostopom (Academic licences):
  - CLARIN.SI licence
  - MULTEXT-East licence
  - Potrebna prijava in posredovanje dodatnih informacij (npr., država, email, namen uporabe, itd.)
- Če uporabnik ne ve, kakšno licenco bi izbral, lahko uporabi Izbirnik licenc "OPEN Licence Selector"

OPEN Licence Selector

Izbirnik

- Izbirnik uporabniku zastavi vprašanja o naravi podatkov in zahtevah dostopa
- Na podlagi odgovorov mu ponudi najprimernejšo (med tistimi, ki jih ponuja repozitorij)

Vmesnik, iskanje in repozitorijski vnosi

Zakaj uporabljati repozitorij CLARIN.SI?

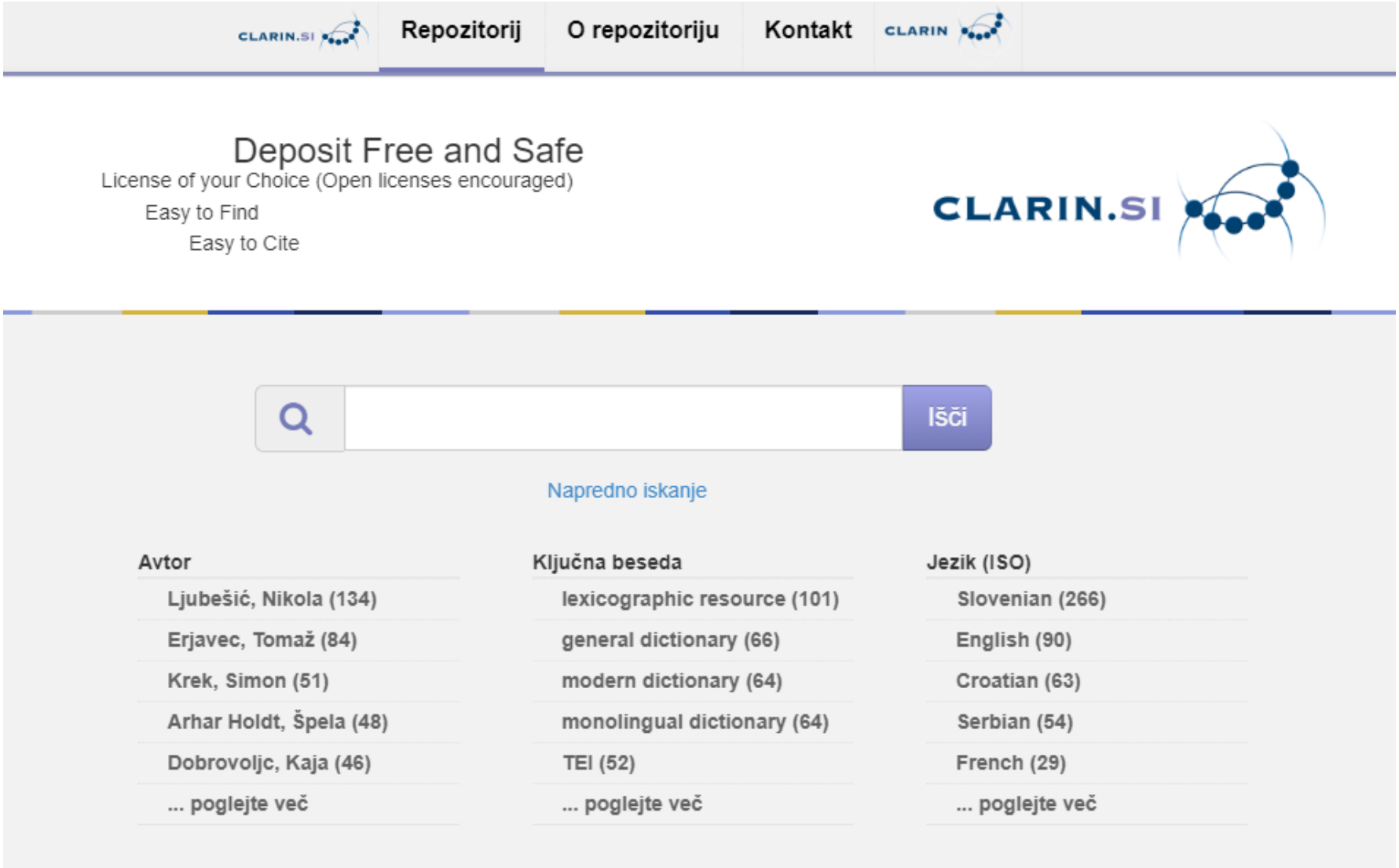
Prevzem

- Prost dostop do širokega nabora podatkov za (predvsem) slovenski jezik
- Kvalitetni in preverjeni (meta)podatki
- Uporabniku prijazen vmesnik in iskalnik

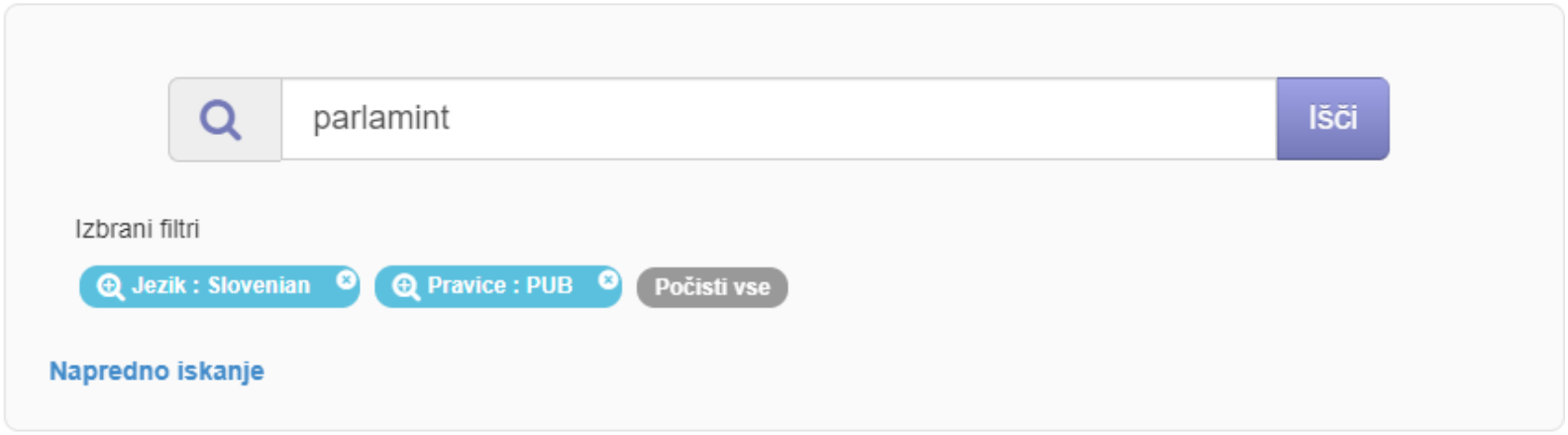
Deponiranje

- Zaupanja vreden repozitorij (CTS)
- Večja možnost citiranja podatkov
- Uredniška podpora

Repozitorij CLARIN.SI - Vmesnik in vnosi



Iskanje in brskanje



Repozitorijski vnos 1/3 - enostavni izpis

Multilingual comparable corpora of parliamentary debates ParlaMint 2.1

Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko: BIBTEX CMDI

Erjavec, Tomaž; et al., 2021, *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1432.

Delite:

CLARIN.SI Data & Tools

Avtorji	Erjavec, Tomaž ; et al. ▶ prikaži vse
Identifikator vnosa	http://hdl.handle.net/11356/1432
URL projekta	https://www.clarin.eu/content/parlamint
Demo URL	https://github.com/clarin-eric/ParlaMint/
Dokumentirano v	https://doi.org/10.1007/s10579-021-09574-0
Datum objave	2021-06-18

Repozitorijski vnos 2/3 - podatki vnosa

Prikaži polni zapis vnosa

Datoteke v tem vnosu

Prenesi navodila za ukazno vrstico

To je vnos Publicly Available z licenco:
Creative Commons - Attribution 4.0 International (CC BY 4.0)

Ime	ParlaMint-BE.tgz
Velikost	144.77 MB
Format	Neznano
Opis	Belgian corpus
MD5	d3cc1f59db6d11c39abd3b0b460e115f

Prenesi datoteko

Ime	ParlaMint-BG.tgz
Velikost	103.13 MB
Format	Neznano
Opis	Bulgarian corpus
MD5	ccf6248f4f3cdf4b5e23d84699c5ee37

Prenesi datoteko

Repozitorijski vnos 3/3 - polni izpis

dc.language.iso	lav
dc.language.iso	lit
dc.publisher	CLARIN ERIC
dc.relation.isreferencedby	https://doi.org/10.1007/s10579-021-09574-0
dc.relation.replaces	http://hdl.handle.net/11356/1388
dc.rights	Creative Commons - Attribution 4.0 International (CC BY 4.0)
dc.rights.uri	https://creativecommons.org/licenses/by/4.0/
dc.rights.label	PUB
dc.source.uri	https://www.clarin.eu/content/parlamint
dc.subject	parliamentary debates
dc.subject	COVID-19
dc.subject	TEI
dc.subject	Parla-CLARIN

Praktični del

Demonstracija

- Demonstracija testnega vnosa - primer slovenskega parlamentarnega korpusa SlovParl 3.0 (1990-1992)
  - Pripravljen izključno za demonstracijo
  - Temelji na starejšem korpusu SlovParl 2.0 (1990-1992)
  - Korpus parlamentarnih razprav Družbeno-političnega zbora Skupščine Republike Slovenije
  - Pokriva obdobje pred, med in po osamosvojitvi Slovenije
  - Izboljšano kodiranje
  - Demo URL
  - Članek (slv)
  - Članek (eng)
- Prikaz repozitorija

Parlamentarni korpus SlovParl 3.0 - Zapis seje (HTML, originalni podatki)

Vsebina zapisa seje

DRUŽBENOPOLITIČNI ZBOR

SKUPŠČINE REPUBLIKE SLOVENIJE

1. seja

(7. maj 1990)

Sejo je vodil dr. France Bučar
Seja se je pričela ob 14.20 uri

PREDSEDUJOČI DR. FRANCE BUČAR: Pričenjam 1. sejo Družbenopolitičnega zbora Skupščine Republike Slovenije, ki je po določbi prvega odstavka 2. člena Poslovnika Družbenopolitičnega zbora Skupščine Republike Slovenije sklicala predsednica zbora, ki je doslej opravljala to funkcijo. Dovolite, da se vam predstavim. Sem France Bučar, kot najstarejši član tega zbora po drugem odstavku 2. člena Družbenopolitičnega zbora Republike Slovenije, zato vodim sejo do izvolitve predsednika zbora. Pri vodenju seje zbora mi bo pomagala dosedanja sekretarka zbora Tina Bitenc-Pengov.
Pred nadaljevanjem našega dela sprašujem delegate v zboru, ali so svojo prisotnost javili službi zbora. Menim, da smo sedaj to ugotovili in bi za vsak slučaj še enkrat pregledali in oddali potrdila o izvolitvi. Menim, da je to sedaj opravljeno. Če kdo tega še ni storil, naj to potem stori.
Na seji je navzočih vseh 80 delegatov tega zbora. Zato prehajamo na predlog dnevnega reda, ki ste ga dobili, in sicer:

Parlamentarni korpus SlovParl 3.0 - Zapis seje v TEI XML (format podatkov v korpusu)

```
<text>
   <front>
      <div type="preface">
         <head>DRUŽBENOPOLITIČNI ZBOR SKUPŠČINE REPUBLIKE SLOVENIJE</head>
         <head type="session">1. seja</head>
         <docDate>7. maj 1990</docDate>
         <note type="chairman">Sejo je vodil dr. France Bučar</note>
      </div>
   </front>
   <body>
      <div>
         <note type="time">Seja se je pričela ob 14.20 uri</note>
         <note type="speaker">PREDSEDUJOČI DR. FRANCE BUČAR:</note>
         <u who="#BučarFrance"
            xml:id="DruzPolZb.1990-05-07.s001-01.u1"
            ana="#chair">
            <seg xml:id="DruzPolZb.1990-05-07.s001-01.seg1">Pričenjam 1. sejo
Družbenopolitičnega zbora Skupščine Republike Slovenije, ki je po določbi prvega odstavka 2.
člena Poslovnika Družbenopolitičnega zbora Skupščine Republike Slovenije sklicala predsednica
zbora, ki je doslej opravljala to funkcijo. Dovolite, da se vam predstavim. Sem France Bučar,
kot najstarejši član tega zbora po drugem odstavku 2. člena Družbenopolitičnega zbora Republike
Slovenije, zato vodim sejo do izvolitve predsednika zbora. Pri vodenju seje zbora mi bo pomagala
dosedanja sekretarka zbora Tina Bitenc-Pengov.</seg>
```